Annette Klosa/Marc Kupietz/Harald Lüngen (Mannheim)

Zum Nutzen von Korpusauszeichnungen für die Lexikographie

1. Einleitung

In diesem Beitrag möchten wir zeigen, wie korpusgestützte Lexikographie von verschiedenen Auszeichnungen in dem der Erarbeitung des Wörterbuches zugrunde liegenden Korpus (= Wörterbuchkorpus) profitieren kann. Solche Auszeichnungen betreffen z.B. die Textstruktur (Überschriften, Absätze, Sätze), die Textsorte (Interview, Gedicht, Todesanzeige …), die Zuordnung eines Textes zu einem Sachgebiet (Sport, Politik, Medizin …) oder die Zuordnung eines Textes zu einem Thema (Gesundheitsreform, Arbeitslosigkeit, Dopingskandal …). Die entsprechende Aufbereitung des Wörterbuchkorpus und eine entsprechende Schnittstelle im Korpusanalyse- und -recherchewerkzeug kann in der lexikographischen Praxis verschiedene Arbeitsschritte erleichtern: Bei einer automatisch gesteuerten Vorstrukturierung und Vorauswahl der Belege zu einem Stichwort können aufgrund einer Auswertung von Textstrukturauszeichnungen beispielsweise solche Kontexte ausgeschlossen werden, in denen das Stichwort in Überschriften vorkommt. Dies ist außerdem besonders relevant für eine vollautomatische Auswahl von Belegen, die ohne redaktionelle Auswahl und Bearbeitung in Wortartikel integriert werden sollen. Auf der Suche nach Kontexten, in denen das Stichwort etwa in der Wiedergabe wörtlicher Rede vorkommt, kann die Markierung solcher Textpassagen bzw. schon eine Auszeichnung der Textsorte die Auswahl entsprechender Belege erleichtern. Es ist offensichtlich, dass für die Ermittlung und Beschreibung fachsprachlicher Lesarten eines Stichwortes eine Annotation von Texten oder auch Textpassa-

DOI 10.1515/lexi.2012–0006

gen hinsichtlich ihrer Sachgebietszugehörigkeit nützlich ist. Die Zuordnung von Belegen zu einem Thema ermöglicht schließlich eine gezielte Vorauswahl von Belegen zur Illustration bestimmter Verwendungskontexte eines Stichwortes. Dabei gibt es natürlich nicht nur ein denkbares thematisches Klassifizierungsschema oder nur eine Textsortenhierarchie für ein Korpus. Im Idealfall sollte daher der lexikographische Nutzer seine eigene Taxonomie und seine eigene Klassifikation im jeweiligen Nutzungskontext beisteuern können.

Ausgehend von der lexikographischen Praxis im Projekt *elexiko* (einem allgemeinsprachlichen Onlinewörterbuch des Gegenwartsdeutschen, vgl. www.elexiko.de) wollen wir anhand verschiedener Beispiele zunächst die Probleme einer adäquaten Wortbeschreibung auf Grundlage eines Wörterbuchkorpus ohne annotierte Texteigenschaften schildern. Dabei werden die redaktionelle Auswahl von Belegen als Grundlage der Artikelerarbeitung (vgl. 2.3.1.1) und die Suche nach illustrativen Belegen zur Integration in die Wortartikel (vgl. 2.3.1.2) besprochen. Die automatische Auswahl von Belegen ist Thema in 2.3.2. Den Nutzen von Korpusauszeichnungen zur Ermittlung von Kollokationen und Konstruktionen zeigen die Überlegungen in 2.4. Insgesamt wird so der große Bedarf an vielfältiger Korpusauszeichnung im lexikographischen Kontext deutlich.

Im dritten Abschnitt werden in 3.1 die im Deutschen Referenzkorpus DeReKo (Kupietz et al. 2010, IDS 2012a) korpusseitig vorhandenen Auszeichnungen nach inhaltlich definierten Gruppen vorgestellt und ihre bereits praktizierte oder ihre potenzielle Nutzung in der Lexikographie unter Rückgriff auf die im ersten Kapitel dargestellten Probleme vorgestellt und bewertet. In 3.2 werden exemplarisch für textbezogene, textlinguistische Auszeichnungen wie Register, Genre und Thema die bisherige thematische Erschließung von DeReKo und ihre Problematiken in Rahmen der lexikographischen Nutzung dargestellt und in 3.3 als Alternative ein Szenario einer zukünftigen nutzerseitigen thematischen Erschließung skizziert.

Im abschließenden Ausblick wird zusammengefasst, wie sich die lexikographische Erarbeitung von Wortartikeln (am Beispiel von *elexiko*) durch die Annotation von Texteigenschaften in Korpora (am Beispiel von DeReKo) verbessern kann und wie in Zukunft eine vollständige Nutzbarkeit der vorhandenen Annotationen bei der lexikographischen Recherche gesichert werden soll und darüber hinaus nutzerseitig definierte Auszeichnungen ermöglicht werden sollen.

2. Praktische Beispiele aus der korpusgestützten Lexikographie

2.1 Das Onlinewörterbuch *elexiko*

Mit dem Projekt *elexiko* wird am Mannheimer Institut für Deutsche Sprache (IDS) ein ausschließlich korpusgestützt erarbeitetes Onlinewörterbuch der deutschen Gegenwartssprache realisiert, das unter der Webadresse www.elexiko.de als eines der Wörterbücher im Wörterbuchportal OWID des IDS aufrufbar ist. Unter einem korpusgestützten Wörterbuch wird dabei verstanden, dass alle lexikographischen Angaben in den Wortartikeln aus dem zugrunde gelegten Wörterbuchkorpus gewonnen werden. In Konsequenz bedeutet dies: „In *elexiko* gibt es keine lexikographische Angabe, die nicht im *elexiko*-Korpus belegt werden kann, es werden keine Angaben aus anderen Wörterbüchern einfach übernommen.“ (Klosa

2011, 16). Hierbei wird das Wörterbuchkorpus sowohl korpusgeleitet (= corpus-driven, also explorativ und ohne Vorannahmen) als auch korpusbasiert (= corpus-based, also ausgehend von einem bestimmten Phänomen und bestimmten Vorannahmen) ausgewertet.[1]

Das *elexiko*-Wörterbuchkorpus umfasst derzeit rund 2,8 Milliarden Textwörter aus 32 Quellen. Es ist als virtuelles Korpus aus DeReKo zusammengestellt und besteht ausschließlich aus Zeitungs- und Zeitschriftentexten.[2] Mit diesem Korpus soll „die Grundgesamtheit der deutschen standardsprachlichen Gemeinsprache in angemessener Weise" gespiegelt werden (Storjohann 2005, 58). Im Laufe der Wörterbucharbeit haben verschiedene Korpusbefunde außerdem zu wichtigen konzeptionellen Ergänzungen in einzelnen Angabebereichen (z.B. den sinnverwandten Wörtern) geführt oder dazu, über die ursprüngliche lexikographische Konzeption (z.B. der typischen Verwendungen) erneut nachzudenken. Hierbei profitiert das Projekt vor allem vom Verfahren der statistischen Kollokationsanalyse (implementiert in COSMAS II; vgl. Belica 1995) und den Recherchemöglichkeiten der Kookkurrenzdatenbank CCDB (vgl. Belica 2001–2007, Keibel/Belica 2007).

In den folgenden Abschnitten soll anhand einiger problematischer Fälle aus der Arbeit mit einem Wörterbuchkorpus darüber nachgedacht werden, wie die lexikographische Praxis weiter optimiert werden könnte, wenn das Korpus in größerem Detail ausgezeichnet wäre. Dabei werden konkrete Beispiele aus der Arbeit am Onlinewörterbuch *elexiko* herangezogen.

2.2 Stichwortliste

Die *elexiko*-Stichwortliste ist vollständig neu und auf der Basis von Korpora erarbeitet worden, und zwar, ohne auf in anderen Wörterbüchern vorhandenen Stichwortlisten aufzubauen (vgl. Schnörch 2005). Sie ist, was für ein im Aufbau befindliches Onlinewörterbuch angemessen erscheint, zugleich dynamisch angelegt, sodass nach wie vor Stichwörter im Rahmen der Lemmatisierungsrichtlinien ergänzt oder gestrichen werden können. Auch die Richtlinien für den Stichwortansatz selbst können im Rahmen des dynamischen Ansatzes prinzipiell noch ergänzt werden, was z.B. im Laufe des Projektes zu einer nachträglichen Festlegung der Lemmatisierung von Pronomina, Artikeln und adjektivischen Sonderformen geführt hat (vgl. Klosa 2011, 166ff.).

Die Stichwortliste wurde in zwei Teilschritten erstellt, wobei hier besonders der erste, automatische Teilschritt interessiert, der korpusgeleitet erfolgte.[3] Bei diesem wurden zwei Verfahren kombiniert: Zum einen die automatische Erstellung einer Stichwortkandidatenliste aus den Korpora, zum anderen ein der Korrektur und Sicherung dienender Abgleich der Stichwortkandidatenliste mit Listen anderer Wörterbücher. Im Umfang wurde die

---

1 Einen detaillierten Einblick in die praktischen Erfahrungen mit diesem Ansatz geben die Beiträge in Klosa (2011). Vgl. zu diesen Ansätzen auch den Beitrag von Petra Storjohann in diesem Band.

2 Auch bei der Zusammenstellung eines Wörterbuchkorpus für den jeweils angestrebten Beschreibungsgegenstand können Auszeichnungen eine wichtige Rolle spielen. So wird z.B. beim *elexiko*-Korpus auf ein den jeweiligen Sprecherzahlen angemessenes Verhältnis von österreichischen, Schweizer und bundesdeutschen Quellen geachtet (vgl. Klosa 2011, 13f.).

3 Zum zweiten, ‚manuell' zu bewältigenden Teilschritt und zu allem Weiteren vgl. ausführlicher Schnörch (2005, 77ff.).

Liste schließlich durch die Anwendung eines auf Frequenz beruhenden Auswahlkriteriums beschränkt. Mithilfe des COSMAS-Lemmatisierers

> „wurden die Flexions- bzw. Paradigmenformen von Wörtern aus den Texten der IDS-Korpora geschriebener Gegenwartsprache Wortformen zugewiesen, die gemäß der Annahmen des automatischen Lemmatisierers als 'wörterbuchübliche Grund- oder Nennformen' gedeutet werden können, z.B. dem Infinitiv bei Verben, dem Nominativ Singular bei Nomen, dem Positiv bei Adjektiven usw." (Schnörch 2005, 75).

Dass die Annahmen des Lemmatisieres darüber, was eine Nennform ist, nicht immer mit den lexikographischen Annahmen hierzu übereinstimmen, ist ebenso zu erwarten wie, dass die automatische Lemmatisierung nicht immer fehlerfrei funktionieren kann. Da die Stichwortkandidatenliste für *elexiko* zu einem Zeitpunkt ermittelt wurde, in der die Korpustexte noch nicht nach Wortarten getaggt waren (vgl. 3.1.2), kam es beispielsweise bei homonymen Wortformen zu falschen Zuordnungen, z.B. *All*: Substantiv, aber auch großgeschriebenes Pronomen am Satzanfang; *behindert*: Adjektiv, aber auch Partizip Perfekt des Verbs. Wäre die *elexiko*-Stichwortliste auf der Basis von nach Wortarten ausgezeichneten Korpustexten erfolgt, hätten einige solcher Fehler leichter erkannt werden können. So konnten sie nur mithilfe der Durchsicht der ganzen Liste durch die Lexikographen gefunden und verbessert werden.

Aus den Stichwortkandidaten wurden die Eigennamen nicht aussortiert. Eigennamen waren – mangels entsprechender Auszeichnung in den Korpustexten zum damaligen Zeitpunkt – generell wie Gattungsbezeichnungen auf ihre jeweilige Grundform zurückgeführt und den Lexikographen zur Prüfung zur Verfügung gestellt worden. Dies hatte den Vorteil, dass Lexeme, die sowohl Gattungsbezeichnung wie Eigenname sind (z.B. *Fischer*, *Hirsch*), nicht als Stichwörter verloren gehen konnten. Andererseits bedeutete dieses Vorgehen aber auch, dass im zweiten, 'manuellen' Schritt der Prüfung der Stichwortkandidaten die Namen als solche markiert werden mussten, um sie gegebenenfalls später aus dem eigentlichen Stichwortbestand des Wörterbuches aussortieren zu können. Die Frage danach, ob Eigennamen in allgemeinsprachlichen Wörterbüchern erfasst und beschrieben werden sollen, wird nämlich sehr unterschiedlich beantwortet (vgl. Klosa/Schoolaert 2011, 196ff.). Für *elexiko* wurde schließlich entschieden, die Eigennamen nicht auszuschließen, sie aber mit einem anderen Angabeninventar zu beschreiben als den appellativischen Wortschatz. Eine Auszeichnung der Namen in den Korpustexten könnte die Lexikographen nun beispielsweise bei einer gezielten Belegauswahl (insbesondere bei homonymen Gattungsbezeichnungen bzw. Eigennamen vom Typ *Fischer*, *Hirsch*) unterstützen (vgl. 2.3.1.2).

Auf der anderen Seite kann sich die Auswahl von Stichwortkandidaten für ein Wörterbuch nicht blind auf automatisch gewonnene Auszeichnungen verlassen, da diese erstens immer fehlerbehaftet sind und zweitens u. U. von linguistischen Vorannahmen (hier zu den Wortarten der jeweiligen Sprache) oder anderen Vorannahmen (z.B., dass nur eine mögliche Interpretation selektiert werden soll) ausgehen, die den Blick auf mögliche Stichwortkandidaten verstellen können. Ein Beispiel, bei dem dies im Fall von *elexiko* leicht passieren könnte, sind Partizipien, die durch Konversion zu Adjektiven werden. Als solche hätten sie dann die Berechtigung, als Stichwörter ins Wörterbuch aufgenommen zu werden. Solche partizipialen Adjektive können lemmatisiert werden[4], wenn sie z.B. die syntaktischen Eigen-

4 Vgl. hierzu genauer Erb (2005, 93f.).

schaften von Adjektiven haben (flektierbar; steigerbar; attributive, prädikative und adverbiale Verwendung), wenn sie in Wortbildungen eingehen (z.B. in Negationen mit *un-*, als Grundwort in Komposita) und insbesondere, wenn sie die ursprüngliche verbale Bedeutung verloren haben (z.B. *bedeutend*, *gewohnt*). Die lexikographische Prüfung der einzelnen Stichwortkandidaten an den Korpusbefunden führt schließlich zu einer fundierten Entscheidung darüber, ob das einzelne Lexem Stichwort wird oder nicht. Damit Kandidaten nicht fälschlicherweise aufgrund von Korpusauszeichnungen ausgeschlossen werden (sogenannte Falsch-Negative), ist eine Vorgehensweise zu wählen, die im ersten Schritt das Auffinden aller Kandidaten (100% Rücklauf) garantiert, aus denen dann – in der Regel iterativ und manuell oder semi-automatisch – alle fälschlicherweise enthaltenen Kandidaten (sogenannte Falsch-Positive) aussortiert werden müssen (vgl. Belica et al. 2011, 466f.). Im Beispiel der Kandidatenliste für Adjektive, könnte es also – je nach verwendetem Tagger – notwendig sein, mindestens auch solche Kandidaten in Betracht zu ziehen, die als Partizip ausgezeichnet sind.

Um die Gefahr Falsch-Negativer zu verringern, können grundsätzlich auch konkurrierende Mehrfachauszeichnungen hilfreich sein. D.h. zum Beispiel, dass zu einer Wortform nicht nur eine Wortart annotiert ist, sondern alle möglichen Interpretationen, die von einem oder besser von mehreren unabhängigen Taggern geliefert werden (vgl. Perkuhn/Keibel/Kupietz 2012, 62f.).

2.3 Belege

2.3.1 Redaktionelle Auswahl von Belegen

2.3.1.1 Belegsuche als Grundlage der Artikelerarbeitung

Die Erarbeitung von Wörterbucheinträgen setzt die möglichst umfassende und sorgfältige Analyse möglichst vieler Belegkontexte voraus. Natürlich variiert die Menge der Textvorkommen eines Stichwortes im Wörterbuchkorpus, was z.B. mit der Wortart des Lemmas zusammenhängt. Die Qualität der Textbelege, die für die lexikographische Bearbeitung zur Analyse zur Verfügung stehen, hängt nicht nur von der Korpuszusammensetzung, sondern auch von den Möglichkeiten, die die Korpusabfrage bietet, ab. Je gezielter ausschließlich Belege zu dem zu bearbeitenden Stichwort ermittelt werden können, desto zutreffender, vollständiger, gründlicher und ergebnisreicher kann die lexikographische Beschreibung werden.

Generell sollten aus der Treffermenge zu einem Suchwort (= Stichwort im Wörterbuch) solche Kontexte ausgeschlossen werden, in denen das gesuchte Lexem in identischen Texten (vgl. 3.1.6) vorkommt (bei einem Zeitungskorpus etwa bei dpa-Meldungen häufig der Fall), da solche Duplikate die Angaben zur Häufigkeit verfälschen. Eine saubere Auszeichnung der Korpus- und Textstruktur (vgl. 3.1.1) ermöglicht es außerdem, solche Kontexte auszuschließen, in denen das Suchwort z.B. in einer Überschrift vorkommt, damit sichergestellt wird, dass nur Bedeutung und Verwendung eines Stichwortes in vollständigen Sätzen untersucht und beschrieben werden. Bibliographische Metadaten wie die Auszeichnung des Erscheinungsortes der Quelle (vgl. 3.1.3) erlauben es nicht nur, das Wörterbuchkorpus hinsichtlich einer den Sprecherzahlen angemessenen Verteilung auf die einzelnen deutschen Sprachräume zusammenzustellen, sondern auch, sich bei der Erarbeitung eines Wort-

artikels zunächst z.B. nur auf die österreichischen Kontexte konzentrieren zu können und anschließend andere Sprachräume zu betrachten, damit gegebenenfalls regionale Varianten in Bedeutung und Verwendung des Stichwortes deutlich hervortreten. Hilfreich wäre darüber hinaus, solche Kontexte von vornherein aus- oder einschließen zu können, in denen das Stichwort in idiomatischer Verwendung oder einer festen Wortverbindung vorkommt. Beispielsweise ist *Abend* im *elexiko*-Korpus sehr häufig in der Verbindung *Heiliger Abend* belegt, sodass etwa die Liste aller Kookkurrenzpartner zu *Abend* das Adjektiv *heilig* in all seinen Flexionsformen sowie Kollokatoren zu dieser festen Verbindung enthält, wodurch andere Kollokatoren zurückgedrängt werden. Der Ausschluss des Vorkommens von *Abend* in der Verbindung *Heiliger Abend* bei der Korpussuche wäre in diesem Fall nützlich, um eine breitere Varianz bei den eigentlichen Kollokatoren zum Stichwort *Abend* ermitteln zu können.[5]

Ein typisches Problem, bei dem die entsprechende Korpusauszeichnung Abhilfe schaffen kann, sind die oben bereits genannten Fälle, bei denen Lexeme wie *Hirsch* oder *Fischer* sowohl als Eigenname wie als Gattungsprädikator belegt sind. Sucht man beispielsweise im *elexiko*-Korpus das Substantiv (*der*) *Stich*, werden zahlreiche Kontexte gefunden, in denen über den Tennisprofi Michael Stich berichtet wird. Eine Kookkurrenzanalyse zu *Stich* bzw. all seinen Flexionsformen erbringt eine Vielzahl von Kollokatoren, die nichts mit dem Gattungsprädikator (*der*) *Stich*, sondern mit dem Familiennamen *Stich* bzw. genauer mit Michael Stich zu tun haben, z.B. *Michael*, *Becker*, *Boris*, *Edberg*, *Elmshorn*, *Wimbledonsieger*. Ohne die Möglichkeit, Kontexte, in denen ein Lexem wie *Stich* als Eigenname vorkommt, mithilfe entsprechender Korpusauszeichnungen auszuschließen, müssen die Lexikographen selbst z.B. mithilfe der Kollokatoren solche Kontexte identifizieren und über Ausschlussformulierungen bei der Suchanfrage die Treffermenge reduzieren, was einen nicht zu vernachlässigenden Zeitaufwand bedeutet. Umgekehrt kann es gerade nützlich sein, nur solche Kontexte im Wörterbuchkorpus zu finden, in denen z.B. *Stich* als Eigenname verwendet wird, etwa, wenn ein Beleg im Wortartikel die Verwendung als Familienname verdeutlichen soll (vgl. Beispiel *Bein* aus *elexiko* in Abbildung 1).

Hilfreich kann außerdem sein, wenn die einem Suchwort zugeordneten Wortformen nur solche Belegkontexte erfassen, in denen das gesuchte Wort in der gewünschten Wortart vorkommt. Sucht man in COSMAS II z.B. alle Kontexte zum Verb *weben*, erfasst der Lemmatisierer auch Kontexte zur Form *web*, in denen nicht der Imperativ Singular (*web*!) des Verbs *weben* belegt ist, sondern englischsprachige Kontexte mit dem Substantiv (*the*) *web* (Beispiel 1) sowie die Verwendung dieser Form in *world wide web* (Beispiel 2) und in Webadressen (Beispiel 3) belegt sind. Um Beispiel 3 als Suchergebnis vermeiden zu können, wäre es sinnvoll, URLs bei der Suche ausschließen zu können.

Beispiel 1: Schäuble sagte, mit dem Projekt «check the **web»** habe sich Deutschland auch international gut positioniert. (dpa, 26.10.2007, (Zusammenfassung 1515), Deutsch für Islamistenpropaganda im Internet immer wichtiger.)

Beispiel 2: Konsequenzen hin oder her, das Interview dauerhaft aus dem Internet zu entfernen, dürfte sich als schwierig erweisen. Hat doch das world wide **web** bekanntlich ein Gedächtnis wie ein Elefant. (Burgenländische Volkszeitung, 23.09.2009, S. 13.)

---

5 Als Behelf kann man eine bestimmte feste Verbindung ausschließen, indem man die Suchanfrage für alle ihre flektierten Formen händisch entsprechend formuliert.

Abbildung 1: Erste Bildschirmseite zum *elexiko*-Wortartikel *Bein* mit Beleg für den Familiennamen *Bein*

Beispiel 3: Auf der Internet-Seite http://szenewob.**web**.my-ct.de/lms können sich die Fans auf einer Deutschlandkarte eintragen, um eine Übersicht über die auswärtigen Fans zu bekommen. (Braunschweiger Zeitung, 05.10.2006, Neues Tippspiel für die VfL-Anhänger.)

Ein ähnliches Problem wie bei den Belegstellen für das Verb *weben* stellt sich bei Homonymen verschiedener Wortarten, z.B. *sieben* als Zahlwort und als Verb. Sucht man Belegkontexte mit dem Infinitiv *sieben* für den Wortartikel zum Verb, werden auch, so lange sich die Suche nicht auf eine Auszeichnung nach Wortarten stützen kann, zahlreiche Kontexte mit dem Zahlwort *sieben* gefunden.

2.3.1.2 Suche nach illustrativen Belegen zur Integration in die Wortartikel

Wie oben erwähnt, werden für die Erarbeitung der Wortartikel in *elexiko* die korpusgeleitete und korpusbasierte Vorgehensweise kombiniert. Dies bedeutet, dass bei bestimmten Arbeitsschritten das Korpus gezielt auf der Suche nach Belegen befragt wird, z.B. um die Bedeutungserläuterung oder einzelne paradigmatische Partner zu illustrieren. Gerade bei der Bedeutungserläuterung, bei der in *elexiko* bis zu drei Belege gegeben werden, wird versucht, mithilfe der Belege nicht nur die Bedeutung des Stichwortes zu illustrieren, sondern zugleich auch Informationen über seinen Gebrauch zu vermitteln, wobei es z.B. zu zeigen gilt, dass ein Substantiv häufiger im Plural, überwiegend in Objektsposition oder tendenziell mit Adjektivattributen verwendet wird. Morphologische und syntaktische Annotationen können die Belegsuche in solch einem Fall unterstützen und beschleunigen, indem etwa ein Substantiv gezielt entsprechend im Singular oder Plural, in Subjekts- oder Objektsposition, einmal mit Adjektivattribut, einmal mit Präpositionalattribut gesucht werden kann usw. Und auch bei den Belegen für paradigmatische Partner wäre eine Unterstützung der Suche nach Belegen, in denen z.B. zwei synonyme Partnerwörter in gleicher syntaktischer Funktion (als Subjekt, Objekt) vorkommen, hilfreich. Allerdings bleibt der Arbeits- und Zeitaufwand auch bei einer umfassenden Unterstützung durch Recherche- und Analysetools für die Sichtung der Fundstellen insbesondere bei sehr frequenten Stichwörtern erheblich (vgl. Storjohann 2011, 102). Es ist offensichtlich, dass vor allem auch Belege, die grammatische Phänomene illustrieren sollen (z.B. bei Adjektiven deren attributive, prädikative oder adverbiale Verwendung oder deren Steigerungsform), leichter und schneller gefunden werden können (zumindest bei häufig belegten Stichwörtern), wenn die Belegkontexte entsprechend annotiert sind.

Auf einer anderen Art der Auszeichnung, nämlich nach textlinguistischer Klassifikation (vgl. 3.1.4), kann die gezielte Suche nach illustrativen Belegen für Angaben zu den Gebrauchsbesonderheiten eines Stichwortes fußen. In *elexiko* werden unter der Überschrift „Gebrauchsbesonderheiten“ beispielsweise Angaben dazu gemacht, ob das Stichwort in der jeweiligen Lesart (also Einzelbedeutung) insbesondere mit Bezug auf ein bestimmtes Thema verwendet wird oder ob es sich um einen fachsprachlichen Terminus einer bestimmten Domäne handelt. So werden im Wortartikel *Schriftsteller* in der Lesart ‘Autor’ zwei unterschiedliche Thematisierungen (im Kontext von Literatur und im Kontext von Politik, vgl. Abbildung 2) beschrieben. Eine Auszeichnung der Korpustexte nach Textthema kann die Belegsuche zur Illustrierung solcher Angaben unterstützen. Ähnliches gilt für die Suche nach Kontexten, in denen das Stichwort *Ton* etwa in der Lesart ‘Teil des Musiksystems’ verwendet wird und die im *elexiko*-Wortartikel die Angabe des Sachgebietes „Musik“ zu verdeutlichen helfen. Eine Auszeichnung der Korpustexte nach Sachgebieten kann bei einem Stichwort wie *Ton* die lexikographische Bearbeitung auch dahingehend unterstützen, dass über die Häufigkeit des Vorkommens in einzelnen Sachgebieten (hier etwa Musik und Töpferei bzw. bildende Kunst) mithilfe einer durch entsprechende Auszeichnung unterstützten Suche Aufschluss gewonnen werden kann.[6] In *elexiko* ist eine Abschätzung zur Häufigkeit der einzelnen Lesart deshalb wichtig, weil die Lesarten entsprechend (von der

6 In *elexiko* werden (wie sonst z. B. bei Ton[1] und Ton[2] üblich) keine Homonyme angesetzt, sondern unter einem Lemmazeichen werden auch semantisch nicht miteinander verwandte Lesarten beschrieben.

häufigsten zur seltensten) angeordnet werden. Eine Abschätzung zur Häufigkeit der einzelnen Lesart kann aber auch dabei helfen zu entscheiden, welche Lesarten im jeweiligen Wörterbuch überhaupt aufgenommen werden sollen. In einem Lernerwörterbuch könnten z.B. generell fachsprachliche Lesarten ausgeschlossen werden, sodass alle Vorkommen, die z.B. dem Sachgebiet „bildende Kunst“ zugeordnet werden können, aus der lexikographischen Bearbeitung ausgeschlossen würden.

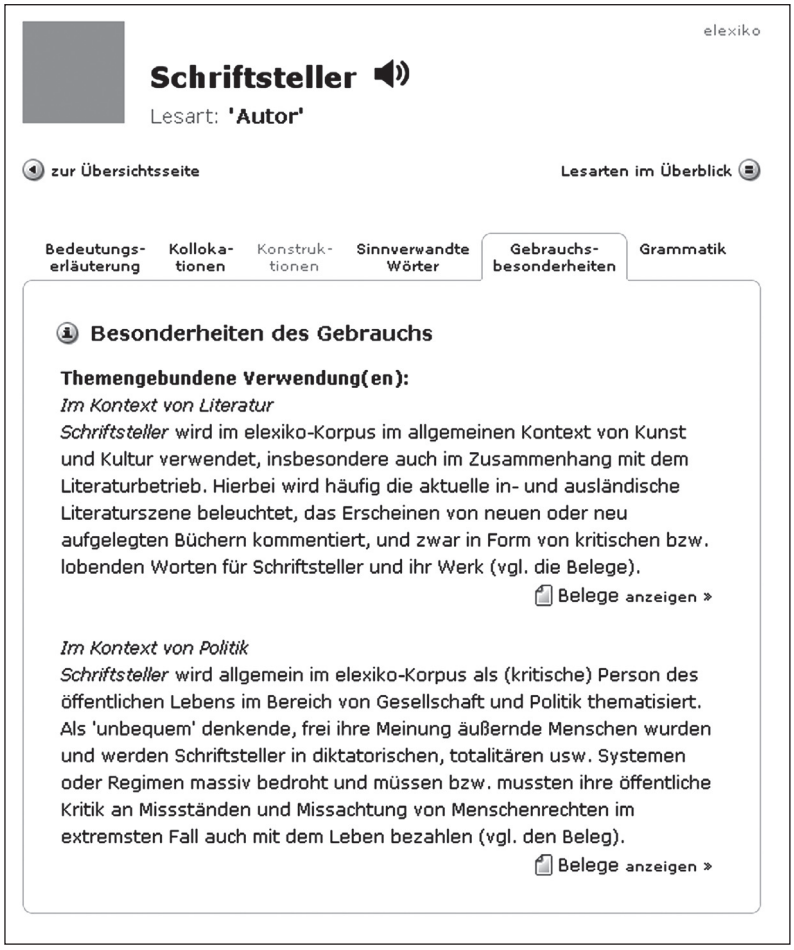

Abbildung 2: Angaben zu Besonderheiten des Gebrauchs im Wortartikel *Schriftsteller*, Lesart ‘Autor’

In vergleichbarer Weise kann eine Auszeichnung nach bestimmten Textsorten bzw. Artikelsorten die Belegsuche für die Angabe der Textsorte in *elexiko* unterstützen, z.B. im Wortartikel *Wasser* in der Lesart ‘Flüssigkeit’ die Angabe zur auffällig häufigen Bindung des Gebrauchs an die Textsorte „Kochrezept“ im *elexiko*-Korpus. (Noch) nicht unterstützt wird hingegen eine Belegsuche, die sich auf Kontexte beschränkt, in denen mündliche Rede wiedergegeben wird oder die einer bestimmten Sprechergruppe (z.B. Jugendlichen) zugewiesen

werden können. Im *elexiko*-Artikel *Alter* in der Lesart ‘männliche Person’ wird beispielsweise unter der Überschrift „Gebrauchsbesonderheiten“ beschrieben, dass dieses Wort häufig als Anrede von Jugendlichen verwendet wird und dass es häufig „in jugendsprachlich und dialektal gefärbter Rede“ belegt ist. Dies illustrieren die Belege in Beispiel 4 und 5.

> **Beispiel 4:** Von 12 Uhr an bildete sich ein dichtes Meer von Luftballons und Transparenten vor einer kleinen Bühne am Neptunbrunnen. Dort rappten eine türkische Jugendband aus Spandau („Mann, **Alter**, stopp die Kürzungen, ey!“) und die Schüler-Kombo „Crossover“. (Berliner Zeitung, 13.03.2000, S. 24 „Mann, Alter, stopp die Kürzungen, ey!“.)

> **Beispiel 5:** Deswegen bin ich gleich zum Chef gestratzt, der residiert da backstage in sonem Hinterzimmer und heißt Natz. Er sacht zu mir: Hey **Alter**, setz dich doch, fährste auch Bock oder was, und ich sach: Na klar Mann! und pflanz mich aufs schwarze Ledersofa hinterm Tisch midder bunten Häkeldecke drauf. (die tageszeitung, 12.04.1999, S. 23, Hell On Wheels.)

Im Idealfall bieten die Korpusauszeichnungen nicht nur ein denkbares thematisches Klassifikationsschema oder nur eine Textsortenhierarchie für ein Korpus, sondern mehrere. Der lexikographische Nutzer sollte diese ergänzen können[7] bzw. dieser auch eine eigene Taxonomie oder seine eigenen Klassifikatoren zur Nutzung hinzufügen können. Dies kann insbesondere relevant werden, wenn ein Wörterbuch zu einem bestimmten Fachwortschatz erarbeitet wird, dessen feinere Untergliederung in Subthemen oder Unterklassen vermutlich (noch) nicht in der allgemeineren Taxonomie bzw. Klassifikation der Korpusauszeichnung enthalten ist.

2.3.2 Automatische Belegauswahl

In *elexiko* werden zu zahlreichen (insbesondere zu wenig frequenten), noch nicht redaktionell bearbeiteten Stichwörtern automatisch aus dem Korpus extrahierte Belege angeboten. Um die Qualität dieser Belege möglichst hoch zu halten, werden die Belege zu einem Stichwort nicht

> „nach rein statistischen Zufallskriterien aus dem Korpus ermittelt, sondern unter Hinzunahme der folgenden, die Auswahl weiter steuernden Kriterien: Die Belege müssen mindestens aus drei verschiedenen Quellen und Jahrgängen stammen. Der Belegumfang soll maximal drei Sätze vor dem Satz mit dem Stichwort und einen Satz danach umfassen, höchstens aber einen Absatz.“ (Klosa 2011, 20).

Außerdem werden, dank entsprechender Auszeichnungen (vgl. 3.1.1 und 3.1.6) solche Kontexte ausgeschlossen, in denen das Stichwort in einer Überschrift erscheint oder in Textduplikaten verwendet wird. Wünschenswert wäre außerdem, wenn Kontexte, in denen Adjektive oder Verben in Substantivierung auftreten (*das Rot*, *das Schlafen*) ausgeschlossen werden könnten, da es sich bei solchen Substantivierungen zumindest teilweise um eigene Stichwörter handelt. Bei Verben wäre es außerdem hilfreich, wenn Kontexte, in denen diese als Partizip I oder II belegt sind, nicht unter den Treffern erscheinen würden, weil diesen unter Umständen ebenfalls eigener Lemmastatus zukommt (s.o.). Beides könnte durch linguistische Annotationen (vgl. 3.1.2) erreicht werden. Die automatische Belegermittlung könnte

---

7 Beispielsweise fehlt in der für DeReKo vorgeschlagenen Klassifikation derzeit ein Themenbereich „(historisches) Adelsleben“ oder überhaupt ein übergeordneter Themenbereich zu „Geschichte“. Vgl. hierzu genauer unter 2.4.

weiterhin dadurch optimiert werden, dass z.B. wo immer möglich mithilfe bibliographischer Metadaten (vgl. 3.1.3) je ein Beleg aus Deutschland, Österreich und der Schweiz ausgewählt wird oder dass wo immer möglich mithilfe textlinguistischer Klassifikationen (vgl. 3.1.4) Belege aus verschiedenen Sachgebieten oder Themen gezeigt werden. Bei der Belegauswahl sollten schließlich mithilfe der Auszeichnung der hierarchischen Korpus- und Textstruktur (vgl. 3.1.1) solche Kontexte ausgeschlossen werden, bei denen das Stichwort in Datumsangaben, Listen, Tabellen usw. oder als Autorenname oder -kürzel vorkommt.

Selbst wenn für ein Wörterbuch auf die Integration automatisch ausgewählter Belege verzichtet wird, kann eine nach den genannten Kriterien gesteuerte Belegauswahl aus dem Korpus genutzt werden. So können diese Kriterien dafür eingesetzt werden, für die lexikographische Bearbeitung nur qualitativ hochwertige, als Belegzitat geeignete Korpusbeispiele zur Verfügung zu stellen. Wenn für die Erarbeitung der lexikographischen Angaben nur mit in diesem Sinne „guten" Belegen gearbeitet wird, ist dies allerdings kein korpusgeleitetes, sondern ein korpusbasiertes Vorgehen. In *elexiko* wird, falls dies bei sehr frequenten Stichwörtern nötig ist, die Treffermenge nicht nach solchen qualitativen Kriterien, sondern ausschließlich nach statistischen Zufallskriterien reduziert.

2.4 Kollokationen und Konstruktionen

In den meisten allgemeinsprachlichen Wörterbüchern werden Kollokatoren z.B. bei den sogenannten Beispielen genannt. In *elexiko* sind sie in einem eigenen Angabebereich unter der Überschrift „Kollokationen" zusammengefasst und dort in Sets aus Fragen und zugehörigen Antworten (in Form von Wortlisten) geordnet. Hinsichtlich der streng korpusgeleiteten Ermittlung dieser Angaben, ihrer Präsentation und ihres Umfangs ist diese Angabe ein „lexikographisches Novum" (Klosa/Storjohann 2011, 49). Bei der Angabe der Kollokationen (oder auch lexikalischen Mitspieler) in *elexiko*

> „repräsentieren die Fragen die Slots, also die Leerstellen in semantisch-konzeptuellen Repräsentationen. Die auf die Fragen antwortenden Wörter stehen für die konkreten Filler oder Partizipanten." (Klosa/Storjohann 2011, 50)

Das Informationspotenzial dieser Angaben hängt dabei sehr stark von der Zusammensetzung des zugrunde liegenden Wörterbuchkorpus ab. Manchmal wirkt sich die Zusammenstellung aus Zeitschriften- und Zeitungstexten negativ aus, weil in diesen Texten bestimmte Themen und Diskurse sehr stark dominieren. So liefert die Kookkurrenzanalyse der Vorkommen des Verbs *antworten* als Bezeichnungen für die Personen, die auf etwas antworten, solche Kollokatoren wie *Experte*, *Kanzler*, *Manager*, *Minister*, *Politiker* usw. Natürlich antworten aber nicht nur Politiker auf eine Frage, sondern im „normalen" Leben antworten alle möglichen Personen auf alle möglichen Fragen. Einer solchen Dominanz der politischen Berichterstattung in den Korpusquellen könnte entgegengewirkt werden, wenn die entsprechenden Texte aufgrund einer textlinguistischen Auszeichnung (vgl. 3.1.4) in einem zweiten Korpusanalyseschritt ausgeschlossen werden können.[8]

---

8 Zum Umgang mit solchen Problemen in der praktischen Artikelarbeit in *elexiko* vgl. genauer Klosa/Storjohann (2011).

Ähnliches wäre dann wünschenswert, wenn die zunächst korpusgeleitete Auswertung der Befunde ergibt, dass bestimmte Stereotype in den Korpustexten transportiert werden, die aber nicht unreflektiert in die Wortartikel eingehen sollen. Ein Vergleich der Kookkurrenzpartner der Wörter *Mutter* und *Vater* ergibt, dass die adjektivischen Kollokatoren von *Mutter* wesentlich häufiger auf den Familienstand oder den Gefühlszustand bzw. das Wesen der Frauen referieren (z.B. *alleinerziehend, alleinstehend, berufstätig; besorgt, depressiv, dominant, fürsorglich, herzensgut, überfordert*) als diejenigen von *Vater*. Umgekehrt werden Väter durch verschiedene Adjektivattribute als tendenzielle Täter gekennzeichnet (z.B. durch *despotisch, gewalttätig, prügelnd, tyrannisch*). Mütter werden eher durch ihre Fürsorge charakterisiert, was sich in verbalen Kollokatoren von *Mutter* wie *aufziehen, großziehen, (sich) kümmern* oder *umsorgen* ausdrückt, Väter dagegen dadurch, dass sie *arbeiten, erzählen, erziehen* oder *spielen*. Auch hier gilt, dass eine thematische Auszeichnung des Korpus und/oder eine Auszeichnung der Textsorte bei der lexikographischen Arbeit ermöglichen würden, Kontexte mit solcher Art auffälligen Thematisierungen und Diskursen wie z.B. „Kriminalität“ oder „moderne Familienkonstellationen“ bei Bedarf aus der Korpusanalyse auszuschließen.

Neben der Angabe der Kollokatoren enthalten einsprachige, auf Beschreibung von Bedeutung und Verwendung ausgerichtete Wörterbücher auch immer die Angabe syntagmatischer Muster bzw. von Konstruktionen mit dem Stichwort. In *elexiko* werden diese korpusgestützt ermittelt und nach syntaktischen Kriterien redaktionell sortiert online präsentiert.[9] Bei sehr frequenten Lesarten bieten die syntagmatischen Muster, die in der Liste der Kookkurrenzpartner erscheinen, die Grundlage für die Auswahl und Formulierung der typischen Verwendungen in *elexiko*. Anders liegt der Fall bei im Korpus nur eher gering belegten, häufig fachsprachlichen Lesarten. Für diese müssen Kontexte gezielt im Korpus gesucht werden, um aufgrund möglichst vieler Belegstellen durch lexikographische Sichtung und Interpretation die typischen Konstruktionen ermitteln zu können. Bei der Suche nach den entsprechenden Belegen (z.B. für das Stichwort *Graf* in der Lesart ‘mathematische Darstellung’) könnte sich daher eine Auszeichnung des Korpus nach bestimmten Sachgebieten (vgl. 3.1.4) als sehr nützlich erweisen, um aus den gefundenen Belegstellen für das Stichwort solche Konstruktionen wie *regelmäßige Grafen*, *Gleichungen als Grafen darstellen* u.Ä. ermitteln zu können.

Ähnliches gilt, wenn eine thematische Auszeichnung des Korpus es erlaubt, nur nach bestimmten Kontexten zu suchen. So wird es möglich, etwa für die Lesart ‘Regierungssitz’ im Wortartikel *Hof* genügend Kontexte aus dem Themenbereich „(historisches) Adelsleben“ zu ermitteln, in denen das Substantiv *Hof* (in der Lesart ‘Regierungssitz’) gemeinsam mit solchen Wörtern wie *Herrscher*, *Herzog*, *König*, *Kammerjungfer*, *Mätresse* usw. belegt ist. In solchen Kontexten kommt *Hof* dann in Konstruktionen wie *am [z.B. kaiserlichen, königlichen] Hof*, *am Hof eingeführt werden*, *das Leben am Hof* usw. vor.

Daneben gibt es Fälle, bei denen im Rahmen der lexikographischen Bearbeitung eines Wortes gezielt nach bestimmten Konstruktionen im Korpus gesucht wird, z.B. nach allen Mustern, die ein Substantiv mit präpositionalem Anschluss zeigen (z.B. beim Stichwort *Arbeit* in der Lesart ‘Tätigkeit’ die Konstruktionen *Arbeit an etwas*, *Arbeit für jemanden/etwas* und *Arbeit mit jemandem/etwas*). Solche Anfragen würden durch eine Auszeichnung des Korpus mit linguistischen Annotationen (vgl. 3.1.2) erleichtert.

---

9 Vgl. hierzu genauer Möhrs (2011).

2.5 Zwischenfazit

Die oben genannten Beispiele haben den Bedarf an vielfältiger Korpusauszeichnung im lexikographischen Kontext deutlich gemacht. Dabei ist festzuhalten, dass dieser Bedarf sowohl die Makro- wie die Mikrostruktur eines Wörterbuches betreffen kann, und zwar genauer sowohl die Stichwortliste als auch einzelne lexikographische Angaben. Der Bedarf entsteht generell in korpusgestützter Lexikographie, insbesondere aber bei Fragestellungen, bei denen das Wörterbuchkorpus nicht Ausgangspunkt der Beschreibung ist (korpusgeleiteter Ansatz), sondern eher der Rückprüfung dient (korpusbasierter Ansatz). Generell kann die Möglichkeit, durch die Berücksichtigung bestimmter Korpusauszeichnungen die Menge an Belegstellen enger einzugrenzen, insbesondere bei Stichwörtern mit einer hohen Frequenz dabei helfen, z.B. bei der Belegauswahl schneller zu einem guten Ergebnis zu kommen. Es gibt aber auch Fälle, wo solche Auszeichnungen gerade in einem niedriger frequenten Bereich die lexikographische Arbeit sehr gut unterstützen würden.

Festzuhalten ist aber auch, dass nicht alle praktischen Probleme korpusgestützter Lexikographie mithilfe von Korpusauszeichnungen gelöst werden können. So kann etwa die automatische Zuordnung von Belegen zu den verschiedenen Lesarten eines Stichwortes nur teilweise und eher behelfsmäßig durch den Rückgriff auf eine Auszeichnung nach Sachgebieten oder Themen gesteuert werden, da nicht alle Lesarten eines Stichwortes notwendigerweise verschiedenen Sachgebieten oder Themen zugeordnet werden können. Welchem Sachgebiet oder welchem Thema sollten etwa die Lesarten ‘Körperteil’ bzw. ‘Teil eines Kleidungsstückes’ des Stichwortes *Arm* zugeordnet werden?

Schließlich ist auch festzuhalten, dass nicht alle Korpusauszeichnungen, die z.B. in DeReKo angeboten werden oder für DeReKo geplant sind, im lexikographischen Kontext eines auf die Zeitungs- und Zeitschriftensprache konzentrierten Wörterbuches wie *elexiko* relevant sind. So sind etwa bibliographische Metadaten (vgl. 3.1.3) hier weniger wichtig, und auch der direkte Nutzen textspezifischer deskriptiv-statistischer Metadaten (vgl. 3.1.5) für die Lexikographie ist nur teilweise gegeben.

3. Korpusauszeichnungen

Das *elexiko*-Wörterbuchkorpus wurde realisiert als virtuelles Korpus aus dem Archiv des Deutschen Referenzkorpus (DeReKo, vgl. Kupietz et al. 2010, IDS 2012), zu einem Zeitpunkt, zu dem noch keine Textauszeichnungen zur Erstellung von virtuellen Korpora verwendet werden konnten. DeReKo ist die weltweit größte Sammlung deutschsprachiger Textkorpora der Gegenwart und jüngeren Vergangenheit und dient als empirische Basis für die germanistisch-sprachwissenschaftliche Forschung. Im März 2012 umfasste sie insgesamt 5,4 Milliarden laufende Textwörter. Eine der Grundannahmen ihres Designs ist, dass ein Korpus weder allgemein repräsentativ noch allgemein ausgewogen sein kann. Daher ist DeReKo nicht in erster Linie als ein fertig verwendbares Korpus konzipiert, sondern als eine so genannte Ur-Stichprobe (vgl. Kupietz et al. 2010), aus der sich idealerweise der Nutzer u.a. anhand von Textauszeichnungen selbst ein virtuelles Korpus zusammenstellt, das im Hinblick auf seine Fragestellung für die von ihm intendierte Grundgesamtheit repräsentativ ist.

Auf DeReKo kann zu wissenschaftlichen Zwecken über die Korpusrechercheschnittstelle COSMAS II des IDS (IDS 1991–2012) zugegriffen werden. DeReKo ist auf vielfältige Weise mit Textauszeichnungen und Annotationen versehen (vgl. Kupietz et al. 2010, Belica et al. 2011), von denen über COSMAS II zur Zeit nur ein Ausschnitt angesprochen werden kann.[10] Um das Potenzial der DeReKo-Auszeichnungen für die Lexikographie zu verdeutlichen, geben wir an dieser Stelle einen Überblick über vorhandene und geplante Auszeichnungen, d.h., ihre Form und ihre Semantik, ihre Abdeckung und ggf. darüber, wie sie ermittelt und annotiert werden.

3.1 Korpusauszeichnungen in DeReKo

Die in DeReKo vorhandenen Auszeichnungen können wie folgt nach inhaltlichen Gesichtspunkten gruppiert werden:

1. Annotation der hierarchischen Korpus- und Textstruktur;
2. Linguistisches Tagging;
3. Bibliographische Metadaten;
4. Textlinguistische Klassifikation;
5. Textspezifische deskriptiv-statistische Metadaten;
6. Textspezifische Metadaten zu Duplikaten.

Die Auszeichnungen unter 1., 3. und 6. sind realisiert als XML-Auszeichnungen nach dem *Corpus Encoding Standard* (XCES, vgl. Ide/Bonhomme/Romary 2000), der für die IDS-Korpora erweitert wurde (IDS-XCES, vgl. Lüngen/Sperberg-McQueen, erscheint). Die Auszeichnungen unter 2. liegen als XML-Standoff-Annotationen zu den XCES-annotierten Texten vor. Die Auszeichnungen unter 4. und 5. erscheinen teilweise im IDS-XCES-Markup und teilweise in einer DeReKo-Metadaten-Datenbank zu den DeReKo-Texten. Im Rahmen der Lexikographie können diese Auszeichnungen (potenziell oder bereits realisiert in COSMAS II) als Grundlage für die Zusammenstellung von virtuellen Korpora, für die korpusbasierte Erstellung der Stichwortliste, für die Korpusrecherche nach Belegstellen sowie für die Präsentation von Rechercheergebnissen dienen.

3.1.1 Auszeichnung der hierarchischen Korpus- und Textstruktur

Die Auszeichnungen der hierarchischen Korpus- und Textstruktur im IDS-Textmodell (vgl. Lüngen/Sperberg-McQueen, erscheint) sind realisiert als XML-Markup in Form von IDS-XCES, einer IDS-spezifischen Erweiterung des Corpus Encoding Standards XCES. Im IDS-Textmodell ist zunächst die dreistufige Hierarchie der DeReKo-Archivstruktur von *Korpus – Dokument – Text* markiert. Ein Korpus entspricht zum Beispiel einem Jahrgang

10 Da COSMAS II ständig erweitert wird, ist es an dieser Stelle nicht sinnvoll, einen genauen Überblick darüber zu geben, welche Funktionalitäten schon implementiert sind. Wir werden uns aber bemühen, folgende Tabellen aktuell zu halten: http://www.ids-mannheim.de/cosmas2/projekt/referenz/annotationen.html und http://www.ids-mannheim.de/cosmas2/projekt/referenz/textklassifikation.html.

von Ausgaben einer Tageszeitung und enthält mehrere Dokumente. Ein Dokument im Fall einer Tageszeitung entspricht der Zeitungsausgabe eines Tages und enthält mehrere Texte, und ein Text (wiederum im Zeitungsbeispiel) entspricht einem einzelnen Artikel oder Kommentar. Aber nicht nur Zeitungen, alle in DeReKo enthaltenen Texte aller Textarten sind in diesem IDS-Textmodell kodiert. Ein Text in DeReKo ist definiert als eine „relativ unabhängige, kohärente Sequenz von Äußerungen in natürlicher Sprache, die aus natürlichen kommunikativen Situationen hervorgegangen ist“ (Perkuhn et al. 2005). Die unter 3.1.4, 3.1.5 und 3.1.6 besprochenen Auszeichnungen beziehen sich auf die Texteinheiten in DeReKo.

Die textstrukturellen Auszeichnungen eines Textes im IDS-Textmodell dienen einer möglichst getreuen Abbildung von Inhalt und Struktur der Textquelle: Es gibt eine hierarchische Kapitel-/Abschnitts-/Absatzstruktur und die Markierung der zugehörigen Überschriften. Des Weiteren sind Listen, Tabellen, Zitate, URLs, Referenzen und Fußnoten und dergleichen ausgezeichnet, ferner Seitenumbrüche und Seitenzahlen (nach der Vorlage). Buchinhalte sind zusätzlich für die Bereiche der Titelei und des Anhangs ausgezeichnet, mit Elementen für Impressum, Widmung, Inhaltsverzeichnis etc. In Dramen und Debattenprotokollen erscheinen Elemente zur Markierung von Sprechern, Äußerungen und Bühnenanweisungen. Für alle Arten von Texten gibt es außerdem diverse „Inline“-Elemente zur Kennzeichnung von typographisch markierten oder fremdsprachigen Textbereichen. Schließlich wird auch die Satzsegmentierung des Fließtextes im IDS-Textmodell markiert.

Neu aufzunehmende Texte werden zunächst mit dem IDS-XCES-Markup des IDS-Textmodells versehen bzw. über Zwischenrepräsentationen in dieses konvertiert. Das IDS-XCES-Markup dient auch als internes Repräsentationsformat für DeReKo in COSMAS II. Derzeit wird eine TEI-P5-kompatible Dokumentgrammatik für das IDS-Textmodell erstellt, die das IDS-XCES ablösen soll (Lüngen/Sperberg-McQueen, erscheint).

Viele dieser Auszeichnungen dienen dazu oder können dazu dienen, bestimmte Textbereiche für Belegstellen auszuschließen, wie in 2.3.1 und 2.3.2 gefordert (z.B. Überschriften, Tabellen, URLs). Markierungen von wörtlicher Rede (vgl. 2.3.1) sind lediglich im Fall von Dramen und Debattenprotokollen vorhanden, nicht jedoch bei belletristischen Texten und Zeitungstexten.

3.1.2 Linguistische Annotationen

Linguistische Annotationen wie Wortarten-Tagging oder syntaktische oder semantische Annotationen beinhalten immer bereits eine Interpretation der Korpusdaten im Licht einer bestimmten Theorie (eines Kategorieninventars), aber auch eines bestimmten Annotationswerkzeugs (vgl. Belica et al. 2011). Die Strategie bezüglich linguistischer Annotationen in DeReKo besteht daher in der Bereitstellung möglichst vieler konkurrierender Annotationsschichten, ggf. auch zu nur einer sprachwissenschaftlichen Analyseebene wie der Syntax. Derzeit bieten wir für DeReKo Korpusannotationen des TreeTaggers der Universität Stuttgart (Schmid 1994) sowie des Machinese Phrase Taggers von Connexor Oy (Tapanainen/Järvinen 1997) an.

Die Annotationen des TreeTaggers umfassen Tokenisierung und Lemmatisierung sowie ein Wortarten-Tagging (Part-of-Speech/POS-Tagging) der Korpustexte. Die verwendeten POS-Tags entsprechen dem Stuttgart-Tübingen-Tagset (STTS, Schiller et al.

1999)[11] für das Deutsche, das mit 54 Tags sehr umfangreich und detailliert ist und somit fast eine morphosyntaktische Annotation darstellt. Die Annotationen des Machinese Phrase Taggers[12] beinhalten Tokenisierung, Lemmatisierung, Wortarten-Tags, morpholexikalische Merkmale, Chunks[13], grobe syntaktische Tags (z.B. für Prä- und Postmodifikator) sowie Informationen zur Wortstellung innerhalb einer NP. Das umfangreiche Inventar der Tags für die Wortarten, Merkmale und Funktionen ist produktspezifisch, aber gut dokumentiert. In beiden Annotationen sind beispielsweise auch Eigennamen als Unterklasse von Substantiven getaggt. Beide Annotationsebenen sind in DeReKo als XML-Standoff-Annotationen realisiert, d.h., sie befinden sich in separaten Dateien und verweisen mittels Referenzen auf die entsprechenden IDS-XCES-Dateien, die auch die Texte enthalten. Nach dem gleichen Prinzip können jederzeit weitere Annotationsebenen hinzugefügt werden.

Das Wortarten-Tagging (inkl. Eigennamen) beider Annotationsebenen ist seit 2009 über die Korpusrecherche-Software COSMAS II in Suchanfragen ansprechbar. Es kann für die in 2.3.1 skizzierte Filterung von Belegstellen nach Wortarten verwendet werden oder für die in 2.3.2 skizzierte Unterscheidung von Partizipien und von Nominalisierungen von Adjektiven. Es kann sogar für die Filterung von Eigennamen bei der Erstellung der Stichwortliste und der Ermittlung von Belegstellen verwendet werden. Für eine Sortierung oder Filterung der Belegstellen nach der Funktion eines syntaktischen Begleiters (wie Subjekt, Objekt, Präpositionalobjekt, vgl. 2.3.1) könnte mit den Auszeichnungen der Tagger lediglich eine Annäherung erzielt werden, eigentlich bedarf es dafür einer tieferen syntaktischen Annotation, z.B. durch einen Dependenzparser.

3.1.3 Bibliographische Metadaten

Ausführliche bibliographische Metadaten werden ebenfalls im IDS-XCES-Markup dargestellt. Für jeden Text und für jedes Dokument werden Autor bzw. Herausgeber, Titel, ggf. Untertitel, Verlag, Erscheinungsdatum und -ort angegeben. Unter den weiteren bibliographischen Metadaten sind hervorzuheben das Datum der Erstveröffentlichung und die Entstehungszeit, die teilweise vom Erscheinungsdatum abweichen (z.B. im Fall von literarischen Werkausgaben) und insbesondere für Untersuchungen sprachlicher Variation über die Zeit benötigt werden. Die bibliographischen Angaben und die chronologische Sortierung, die von COSMAS II präsentiert werden, beruhen auf diesen Metadaten. In der DeReKo-Metadaten-Datenbank ist zusätzlich die aus dem Erscheinungsort abgeleitete Zuordnung eines Textes zu einer der Regionen D, AT, CH oder DDR kodiert. Über die Regionenzuordnung und den Erscheinungsort können Belege nach bestimmten Regionen sortiert oder gefiltert werden (vgl. 2.3.1) und somit auch regionentypische Lesarten ermittelt werden.

---

11 Übersicht: http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html

12 Übersicht: http://www.ids-mannheim.de/cosmas2/projekt/referenz/connexor/

13 Chunks sind phrasenähnliche Wortgruppen unterhalb der Satzebene, „in denen typischerweise ein Inhaltswort mit umgebenden Funktionswörtern gruppiert ist“ (Abney 1991). Im Gegensatz zu Phrasen sind sie aber nicht rekursiv, d.h., ein Chunk kann nicht Teil eines (größeren) Chunks sein.

3.1.4 Textlinguistische Klassifikation

Die Auszeichnungen dieser Gruppe betreffen textlinguistische Charakterisierungen von Dokumenten oder Texten als Ganzes. Die hier kodierten Dimensionen wie Textart und Textsorte sind nicht orthogonal zueinander, sondern können überlappen oder voneinander abhängen. Mit ihnen können Filterungen von Belegen nach Textsorte, Sachgebiet oder Textthema vorgenommen werden, was wiederum zu einer Filterung nach Lesarten beitragen kann (vgl. 2.3.1 und 2.4).

Textart: eine ursprünglich manuell annotierte Textart aus einem Inventar von derzeit ca. 170 Kategorien (z.B. *Zeitung*, *Roman, Brief, Gesetz, Lehrbuch, Drama, Comic*). Daneben weisen einige Texte zusätzlich die Angabe einer spezifischeren Textart wie *Zeitung:Tageszeitung* auf.

Textsorte (nur Zeitungstexte): Textsorte des Artikels. Wenn auf der Dokumentebene als Textart *Zeitung* angegeben ist, ist für die Textebene eine Artikeltextsorte wie *Bericht, Leserbrief, Interview* markiert.

Ressort (alternative Bezeichnung: „column"; nur bei Zeitungstexten): Sachgebiet von Artikeln wie in der Kopfzeile der Zeitungsseite ursprünglich angegeben: *Politik, Hessen, Hochschule, Immobilien, ...*

Normalisiertes Ressort (nur Zeitungstexte): Sachgebiet eines Artikels aus einem übergreifenden Inventar von Zeitungsressorts: *Politik, Kultur, Sport*, ... Das Ressort *Leibesübungen* der „tageszeitung" wird beispielsweise auf das normalisierte Ressort *Sport* abgebildet.

Textthema (alternative Bezeichnung: „topic"): Jeder Text ist für die zwei wahrscheinlichsten Themen aus einer zweistufigen, allgemeinen Thementaxonomie mit 12 übergeordneten und 52 untergeordneten Themen mit Angabe ihrer Wahrscheinlichkeit annotiert. Beispiele für übergeordnete Themen sind *Politik, Wissenschaft, Gesundheit/Ernährung, Kultur, Natur/Umwelt* und *Freizeit/Unterhaltung*; Beispiele für untergeordnete Themen sind *Sport:Tennis* und *Kultur: Literatur* (für die vollständige Taxonomie vgl. Tabelle 2). Die Themenklassifikation erfolgt automatisch durch einen Naive-Bayes-Klassifikator (vgl. Weiß 2005). Da die lexikographische Nutzung thematischer Klassifikationen von Korpustexten ein praktisches Problem (vgl. 2.3.1 und 2.4) und auch ein aktuelles Forschungsthema darstellt (vgl. 3.3), berichten wir in Abschnitt 3.2 genauer über die Erfahrungen mit der Erstellung der zugrunde liegenden Taxonomie und der Anwendung des Klassifikators.

3.1.5 Textspezifische deskriptiv-statistische Metadaten

Bestimmte textspezifische, deskriptiv-statistische Metadaten werden direkt aus den Korpustexten oder der IDS-XCES-Annotation extrahiert und in unserer DeReKo-Metadaten-Datenbank gespeichert. Sie enthält Felder mit Angaben zur Anzahl der in einem Text enthaltenen Tokens, Wörter, Stoppwörter, Zahlen, Sätze, Absätze, Seitenumbrüche und zur Textlänge (mit dem und ohne den Header-Bereich in IDS-XCES) in Bytes. Ein weiteres Feld kodiert

die geopolitische Zuordnung eines Texts als ‘D’, ‘AT’, ‘CH’ oder ‘DDR’ anhand des Erscheinungsortes. Noch ein weiteres Feld kodiert die Anzahl von Schreibweisen, die der neuen deutschen Rechtschreibung zuzuordnen sind, eines die Anzahl von Schreibweisen, die der alten deutschen Rechtschreibung (von vor 1999, wie ‘daß’) zuzuordnen sind, und eines die Anzahl von Vorkommen nicht-deutscher Zeichen (wie ‘ş’ oder ‘å’). Diese Angaben spielen in der Lexikographie eine geringe Rolle, allenfalls noch die genannte geopolitische Zuordnung, vgl. 2.3.1).

3.1.6 Textspezifische Angaben über Duplikate

Relationale Metadaten über Duplikate oder ähnliche Texte zu einem Text oder über repetitive Texte werden derzeit ebenfalls im IDS-Textmodell (IDS-XCES-Markup) dargestellt. Diese dienen grundsätzlich dazu, Stichprobenverzerrungen, die sich aus unerwünschten Reduplikationen von Textpassagen ergeben können, vermeidbar zu machen. In der Lexikographie dienen sie der Korrektur von Häufigkeitsangaben (vgl. 2.3.1) und der Filterung von Belegen (vgl. 2.3.2). Gründe, warum Ähnlichkeiten ausgezeichnet werden und Dubletten nicht einfach gelöscht werden, sind z.B., dass es oft keine vollständigen Übereinstimmungen gibt, man nicht mit Sicherheit sagen kann, was Original und was Dublette ist, oder welcher aus einem Cluster ähnlicher Texte tatsächlich veröffentlicht wurde, und dass, wenn eher die Rezeption als die Produktion untersucht werden soll, u.U. auch vollständige Dubletten mit in die Stichprobe mit einbezogen werden müssen (vgl. Kupietz 2005).

Die Metadaten über Duplikate sind wie folgt konzipiert: Auf der obersten Auszeichnungsebene gibt es zunächst drei sich ausschließende Hauptkategorien, mit denen Texte annotiert werden: DELETE, VARIANT und REFERENCE.

DELETE markiert Texte, die für eine Weiterverarbeitung etwa durch COSMAS II nicht empfohlen sind, da es sich wahrscheinlich um technisch bedingte Kopien oder um vollständige Kopien handelt, die in sehr kleinen zeitlichen Abständen erzeugt wurden. Eine Subkategorisierung von DELETE markiert diese Texte mit einer von vier „Begründungen“ für diese Klassifikation:

- ‘classified’: der Text gehört zu einem manuell klassifizierten Cluster ähnlicher Texte (s.u.);
- ‘duplicate’: der Text ist identisch mit einem Referenztext;
- ‘version’: der Text gilt als eine Version des Referenztextes (mindestens 90% gemeinsame Pentagramme, d.h. Ketten von fünf aufeinanderfolgenden Tokens, und ein ähnliches Veröffentlichungsdatum);
- ‘repetitive’: der Text ist repetitiv (enthält zu 95% sich wiederholende Pentagramme mit einer Wiederholungsrate von mindestens 10).

VARIANT markiert Texte, zu denen zwar ähnliche Texte in DeReKo existieren (mindestens 70% gemeinsame Pentagramme), die aber die o.g. Bedingungen für die Klasse DELETE nicht erfüllen. Als VARIANT werden solche Texte ausgezeichnet, die wahrscheinlich veröffentlicht wurden, als DELETE:version dagegen solche, die vermutlich nicht veröffentlicht wurden (z.B. fälschlicherweise aus einem Redaktionssystem exportierte Texte). Genauere Informationen geben außerdem die Attribute in Tabelle 1.

Attributname	**Bedeutung**
reference	ID des Referenztextes
coverage	Anteil der Abdeckung von Text und Referenztext durch gemeinsame Pentagramme
unmatched	Anteil der tokens, die nur im Text, aber nicht im Referenztext vorhanden sind
d-day	Abstand in Tagen zwischen Erscheinungsdatum von Text und Referenztext
transpositions	Ungefähre Anzahl der notwendigen Transpositionen von Token-Sequenzen, um den Text dem Referenztext möglichst ähnlich zu machen.
cluster_size	Anzahl der Dokumente, die in der Relation stehen, die in der Subkategorie angegebenen ist
cluster_name	Kurzklassifikation der Textinhalte (nur bei sehr großen, manuell klassifizierten Clustern)
date	Datum der Markierung des Textes mit den duplikatbezogenen Kategorien
max-rep	(nur bei Hauptkategorie 'DELETE' mit Subkategorie 'repetitive':) Maximale Wiederholungsanzahl eines Pentagramms
order	(nur bei Hauptkategorie 'DELETE' mit Subkategorie 'classified':) order=2, falls der Text selbst nicht manuell klassifiziert wurde, aber mindestens neunzigprozentige Ähnlichkeit zu einem Text aus einem manuell klassifizierten Cluster aufweist

Tabelle 1: Weitere Attribute für die Auszeichnung von Textduplikaten

Das folgende sind zwei Beispiele für Textauszeichnungen nach diesen Kategorien:

1. `DELETE:duplicate:[reference="A98/OKT.63396", coverage="1.00", unmatched="0.00", d-day="818", transpositions="0", cluster_size="2", date="2005-02-24"]` (Hintergrund: Eine Schweizer Tageszeitung druckte eine Zeitlang regelmäßig Bibelzitate ab, die außerhalb von Artikeln standen und somit als kurze, eigenständige Texte gelten; in diesem Fall griff man nach 818 Tagen auf ein schon einmal verwendetes Zitat zurück.)
2. `VARIANT:variant:[reference="M08/MAI.41071", coverage="0.73", unmatched="0.06", d-day="1", transpositions="0", cluster_size="2"]` (Hintergrund: Ein Artikel über einen Box-Vizemeister erschien in einer Mannheimer Tageszeitung einmal im regionalen Ressort „Rhein-Neckar" und am Tag darauf in dem lokalen Ressort „Bergstraße" mit abweichender Überschrift.)

Die dritte mögliche Hauptkategorie, REFERENCE, schließlich markiert Texte, auf die von den anderen als Referenztext verwiesen wird.

3.2 Thematische Erschließung

Die thematische Erschließung eines Korpusarchivs wie DeReKo (im Sinne einer Auszeichnung der Korpustexte mit einer thematischen Kategorie) ist eine anspruchsvolle Aufgabe. Zunächst wäre es aus Gründen der Interoperabilität mit anderen Korpora wünschenswert,

könnte für die Annotation ein für alle Arten von Texten anwendbarer, formalisierter und formal referenzierbarer Themenkatalog (eine Thementaxonomie oder -ontologie) für Texte, etwa aus dem Bibliothekswesen, verwendet werden. Andererseits gibt es die Anforderung, dass eine Thementaxonomie datengetrieben sein sollte, d.h., genau die und nur die Kategorien enthalten sollte, die in dem zu klassifizierenden Korpus vorkommen. Bei der thematischen Erschließung von DeReKo wurde versucht, einen Ansatz zu entwickeln, der beide Anforderungen berücksichtigt und im Folgenden anhand Weiß (2005) wiedergegeben wird.

Um eine geeignete Thementaxonomie zu konstruieren, wurde zunächst anhand von rund 400.000 Trainingsdokumenten (Zeitungsartikeln aus den Jahren 1985–2002[14]) ein Dokumentclustering nach dem Verfahren von Karypis/Zhao (2002) durchgeführt. Die entstandenen Cluster (100 pro Jahrgang) wurden einer Qualitätskontrolle unterzogen und dabei inspiziert und manuell mit einem Ad-hoc-Thema wie z.B. „affaere_barschel_87“ annotiert. In einem weiteren Schritt wurden diese Ad-hoc-Themen dann passenden thematischen Kategorien der Themenontologie *Open Directory* (s. Netscape 2012), z.B. „Politik:Inland“ zugeordnet (vgl. Weiß 2005).

Hauptkategorie	**Unterkategorien**
Fiktion	Vermischtes
Freizeit/Unterhaltung	Reisen, Rundfunk, Vereine/Veranstaltungen
Gesundheit/Ernährung	Gesundheit, Ernährung
Kultur	Bildende Kunst, Darstellende Kunst, Film, Literatur, Mode, Musik
Natur/Umwelt	Garten, Tiere, Wetter/Klima
Politik	Ausland, Inland, Kommunalpolitik
Sport	Ballsport, Fußball, Motorsport, Radsport, Tennis, Vermischtes, Wintersport
Staat/Gesellschaft	Arbeit/Beruf, Bildung, Biographien/Interviews, Drittes Reich/Rechtsextremismus, Familie/Geschlecht, Kirche, Recht, Tod, Verbrechen
Technik/Industrie	EDV/Elektronik, Kfz, Transport/Verkehr, Umweltschutz, Unfälle
Wirtschaft	Banken, Bilanzen, Öffentliche Finanzen, Sozialprodukt, Währung
Wissenschaft	Populärwissenschaft
Rest	Börsenkurse, Geburt/Tod/Heirat, Impressum, Inhaltsverzeichnisse, Ligatabellen, Tabellen, Veranstaltungshinweise

Tabelle 2: Aktuelle Taxonomie von Themenkategorien für DeReKo (s. Weiß 2005)

Das Open Directory ist ein Katalog für Webdokumente mit dem Anspruch, alle Themengebiete von Webtexten zu erfassen. Daten und Webseiten können mit Open-Directory-Kategorien verlinkt werden. Es enthielt zum Zeitpunkt der hier beschriebenen Arbeiten (2005) 590.000 thematische Kategorien, von denen allerdings für die DeReKo-Taxonomie nur ein Bruchteil übernommen wurde. Zum einen wurden spezialisierte Themen (ab der dritten Ebene) weggelassen (beispielsweise wurde „Kultur:Film“ übernommen, „Kultur:Film:Filmver-

14 Diese Jahreszahlen sind rekonstruiert.

leih“ aber nicht) und zum anderen wurden eng verwandte Themen, die im Open Directory zu verschiedenen Oberthemen gehören, zu einer Kategorie vereinigt. Die so entstandene Taxonomie für DeReKo enthält 12 Oberkategorien und 52 Unterkategorien, die alle einer Kategorie oder einer Vereinigung von Kategorien des Open Directory entsprechen (siehe Tabelle 2). Durch das oben beschriebene Clustering und die Zuordnungen der Cluster zu den Themenkategorien wurden indirekt auch alle Dokumente des Trainingskorpus einer Themenkategorie aus der Taxonomie zugeordnet; durch Heuristiken wurden noch einige weitere Dokumente für Themen mit geringer Dokumenthäufigkeit ermittelt und hinzugefügt. Das so entstandene Korpus wurde im Folgenden als Trainingskorpus für einen Naive Bayes-Themen-Klassifikator verwendet. Die Performanz dieses Klassifikators lag bei 83% durchschnittlicher Präzision und 82% durchschnittlicher Vollständigkeit auf einem Testkorpus von 30 Zeitungsartikeln je Unterkategorie aus dem Jahre 2003 (vgl. Weiß 2005). Alle DeReKo-Texte wurden und werden seither mit den laut diesem Klassifikator zwei wahrscheinlichsten Themenkategorien annotiert.

Die automatische Themenauszeichnung ist allerdings auf den bestehenden DeReKo-Texten nicht immer so genau, wie die oben beschriebene Evaluation suggeriert. Das liegt offenbar daran, dass sowohl die Taxonomie anhand von Zeitungsartikeln entwickelt als auch das Trainingskorpus aus Zeitungstexten zusammengestellt wurde, DeReKo jedoch mittlerweile auch einen beachtlichen Anteil anderer Textarten enthält. Auch konnte die beschriebene teilautomatisierte Annotation der Trainingsdokumente mit Themenkategorien nicht so präzise ausfallen wie eine manuelle Annotation durch einen Experten – wobei der Clusteringansatz es aber erst ermöglicht hat, ein Trainingskorpus im erforderlichen Umfang herzustellen.

3.3 Nutzerseitige thematische Erschließung

Ein größeres Problem als die mangelnde Genauigkeit der bestehenden automatischen Themenauszeichnung von DeReKo-Texten ist, dass jede Taxonomie nicht aus den beobachteten Daten abgeleitet werden kann und daher in dieser Hinsicht zunächst arbiträr ist. Das heißt, ob eine Taxonomie und ihre Umsetzung besser ist als eine andere, welche Kategorien relevant sind und welche nicht, und wo die Grenzen zwischen diesen liegen sollten, hängt entscheidend vom Verwendungszweck ab (beispielsweise der Erstellung eines allgemeinsprachlichen Wörterbuchs vs. der Erstellung eines Fachwörterbuchs, vgl. 2.3.1). Das gilt im Prinzip gleichermaßen für Textsorten und -genres. Mit der bestehenden thematischen Taxonomie wurde einerseits – wie oben beschrieben – versucht, den Daten möglichst gerecht zu werden und andererseits, sie im Hinblick auf einen durchschnittlichen oder typischen Verwendungszweck zu optimieren. Darüber hinaus wäre es allerdings wünschenswert, auch Textklassifikationen – nicht nur entlang einer thematischen Dimension – zu ermöglichen, die für spezielle Anwendungen optimiert sind.

Gerade aufgrund des Urstichprobendesigns von DeReKo mit der Option der nutzerseitigen Definition von virtuellen Korpora im Hinblick auf bestimmte Forschungsfragen wäre eine solche Möglichkeit besonders relevant. Um die Definition von virtuellen Korpora anhand von Textauszeichnungen optimal bewerkstelligen zu können, wäre es konsequent und besonders wünschenswert, wenn der Nutzer nicht nur von gegebenen Auszeichnungen ausgehen könnte, sondern selbst in der Lage wäre solche hinzuzufügen. Ein Hindernis bei

der Umsetzung eines solchen Vorhabens ist, dass Nutzer nicht etwa alle Texte herunterladen können, um diese lokal bei sich entsprechend ihrer Vorstellungen auszuzeichnen, da die Texte urheberrechtlich geschützt sind und Lizenzverträge typischerweise nur eine stark eingeschränkte Nutzung erlauben, die das Kopieren vollständiger Texte nicht umfasst. Der Weg, den das IDS 2010 mit dem KorAP-Projekt (siehe Bański et al. 2012) zur Lösung dieser grundsätzlichen Problematik eingeschlagen hat, folgt in etwa dem Motto: „wenn die Daten nicht zu den Programmen dürfen, müssen die Programme zu den Daten kommen“.[15] Das heißt, Ziel der Nachfolgeplattform von COSMAS II, die im Rahmen von KorAP entwickelt werden soll, ist es, die Voraussetzungen dafür zu schaffen, dass nutzerdefinierte Programme zur Analyse und zur Annotation in einer kontrollierten Umgebung („Mobile Code Sandbox“) auf DeReKo angewendet werden können. Da speziell für Textklassifikatoren die Anforderungen an eine solche kontrollierte Umgebung gering sind, soll eine Schnittstelle für nutzerdefinierte Textklassifikatoren in einer ersten Stufe, wie im Folgenden beschrieben, bereits 2012 unabhängig von KorAP implementiert werden. Unser Szenario beruht auf Weka (Waikato Environment for Knowledge Analysis, Witten/Frank/Hall 2011), einer Softwareumgebung, in der Klassifikatoren auf relativ einfache Weise realisiert werden können.

Weka ist eine in Java implementierte, unter der GNU General Public License verfügbare Software, welche eine Reihe von Data-Mining-, insbesondere Klassifikations-Algorithmen bietet, die über einheitliche Schnittstellen angesprochen werden können. Weka wird auch in der korpus- und computerlinguistischen Forschung und Lehre zur automatischen Textklassifikation genutzt, vgl. z.B. Dipper (2011). Zwar setzt die Benutzung der Weka-Klassifikationsalgorithmen Kenntnisse über maschinelles Lernen und automatische Klassifikation voraus, sind diese aber vorhanden, ermöglicht die Software die Entwicklung und Evaluation von eigenen Klassifikatoren praktisch ohne Programmierkenntnisse. Das durch Weka eingeführte ARFF-Format für die Vektorraumrepräsentation von Datensätzen ist einheitlich für alle Algorithmen. Die Weka-Algorithmen und -Filter können über eine graphische Benutzerschnittstelle („Explorer“), von Programmen aus über die Java-API oder von der Kommandozeile aus angesprochen werden. Hat man mit Weka einen Klassifikator mit den eigenen Daten inklusive einer eigenen Taxonomie trainiert und evaluiert, kann man diesen als Java-Objekt abspeichern und später in derselben oder in einer anderen Weka-Instanz zur automatischen Klassifizierung weiterer und neuer Daten verwenden („prediction“-Funktionalität). All dies sprach dafür, Weka in unser erstes Szenario der Anwendung nutzerdefinierter Programme auf DeReKo zu integrieren.

Um die Schnittstellen und den Workflow für ein solches Szenario mit Weka zu eruieren, haben wir zunächst den unter 3.2. beschriebenen Klassifikator anhand der Spezifikationen in Weiß (2005) in Weka nachgebaut. Unser Szenario beinhaltet nun, dass Phase I beim Nutzer stattfindet, während Phase II in der DeReKo-Umgebung erfolgt, wie in Abbildung 3 dargestellt.

---

15 Dies ist natürlich nicht nur aus rechtlichen Gründen meistens notwendig, sondern bei großen Korpora auch informatisch sinnvoll.

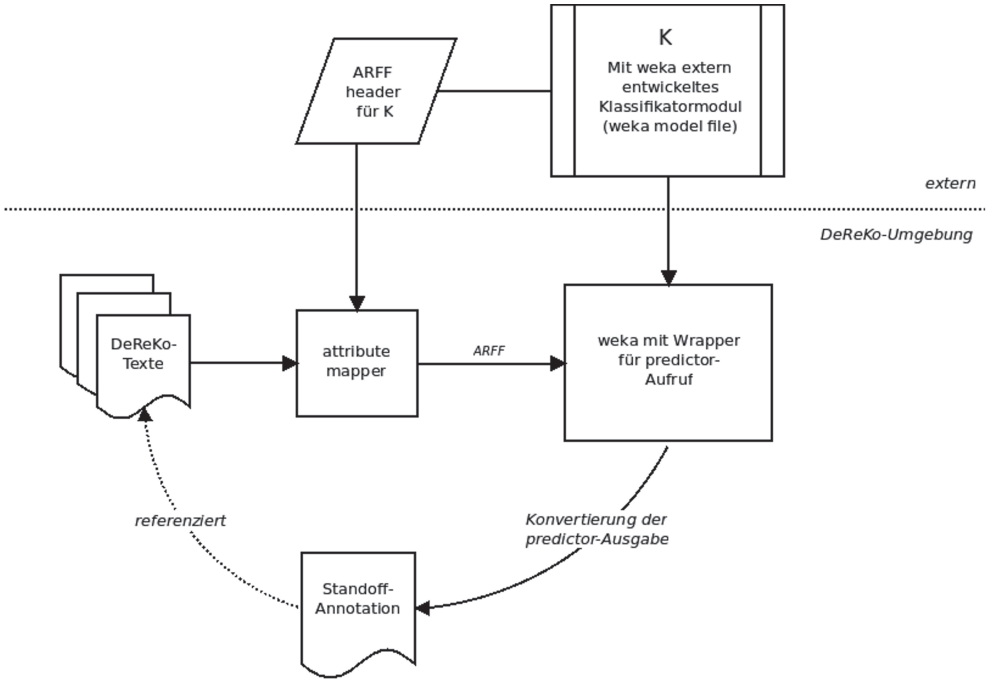

Abbildung 3: Szenario für Anwendung nutzerdefinierter Programme auf DeReKo am Beispiel automatischer Textklassifikation

Ein zu lösendes Problem in diesem Szenario ist, dass die DeReKo-Texte bei der Klassifikation (prediction) in dem gleichen Vektorraum dargestellt sein müssen, der beim (externen) Training des Klassifikators K verwendet wurde, d.h., dass die DeReKo-Texte für die Klassifikation mit K genau mit den Attributen von K dargestellt werden müssen. Dies bedeutet, dass ein DeReKo-Text, der in diesem Ansatz normalerweise durch die in ihm enthaltenen Wortformen dargestellt wird, alternativ anhand der Wortformen, die in den Trainingsdokumenten von K enthalten waren, dargestellt werden muss. Formal sind diese im Header-Abschnitt der ARFF-Datei, die beim Training von K verwendet wurde, spezifiziert. Der ARFF-Header muss also mit dem Modul K an die DeReKo-Umgebung übergeben werden, und die DeReKo-Texte müssen zunächst in einem *attribute mapper* auf die im Header spezifizierte Attributmenge (Menge der Wortformen der Trainingsdokumente von K) abgebildet werden. Dabei kann es natürlich zu Informationsverlusten kommen, daher müssen mit K und dem Header eigentlich noch weitere Informationen übergeben werden. Im Grunde muss für jedes in K verwendete Attribut sein Typ und die Methode seiner Extraktion aus den Daten bekannt sein. Fürs Erste beschränken wir uns aber hier auf Wortformen. Diese können im Vergleich zu Lemmata mehr Klassen unterscheiden, wie Weiß (2005) argumentiert. Aber auch bei Wortformen als Attribute sind Zusatzinformationen erforderlich, wie die, welcher Tokenizer verwendet wurde oder ob Groß- und Kleinschreibung unterschieden wurde. Im einfachen Fall handelt es sich dabei um die Angabe, welche Weka-Filter mit welchen Settings verwendet wurden, die sich auch im ARFF-Header befindet. Im komplizierteren Fall

muss die gesamte Vorverarbeitung der Trainingsdaten von K bekannt gemacht werden. In der Praxis werden seitens der DeReKo-Umgebung Vorgaben gemacht werden über erlaubte Attribute und ihre Extraktion für das Training von K.

In dem Szenario muss der Nutzer auch selbst sicher stellen, dass die Textdaten die er zum Training seines Klassifikators verwendet, möglichst kompatibel sind mit dem DeReKo-Ausschnitt, für den er die thematische Annotation benötigt, beispielsweise, dass es sich um Texte der gleichen Textsorte (wie Zeitungsartikel) handelt. Nur so kann man davon ausgehen, dass die Performanz von K auf den DeReKo-Texten in etwa der Performanz von K auf den Testdaten des Nutzers entspricht. (Wie oben dargestellt, können aus lizenzrechtlichen Gründen leider keine Texte aus DeReKo als Trainingsdaten zur Verfügung gestellt werden).

4. Zusammenfassung und Ausblick

Anhand von Beispielen aus dem Onlinewörterbuch *elexiko* wurde gezeigt, dass vielfältige Korpusauszeichnungen die lexikographische Erarbeitung der Wortartikel erheblich unterstützen können. Dies betrifft zunächst die Schaffung der eigentlichen Arbeitsgrundlage, nämlich des Wörterbuchkorpus, das je nach geplantem Wörterbuchgegenstand mithilfe solcher Korpusauszeichnungen wie z.B. der Textsorte gezielt zusammengestellt werden kann. In einem weiteren Schritt kann die Erstellung einer Stichwortkandidatenliste von Korpusauszeichnungen profitieren, z.B. indem als solche markierte Eigennamen in den Stichwortbestand ein- oder ausgeschlossen werden können. Die Erarbeitung der Wörterbucheinträge schließlich kann in vielfältiger Weise durch Korpusauszeichnungen gewinnen, z.B. bei der automatischen wie der manuellen Ermittlung von Belegen, die zur Illustration in die Wortartikel integriert werden, oder bei der Ermittlung von Kollokationen und Konstruktionen.

Nicht alle Korpusauszeichnungen haben allerdings eine direkte Relevanz im lexikographischen Kontext (z.B. bibliographische Metadaten). Und nicht jedes lexikographische Problem kann mithilfe von Korpusauszeichnungen gelöst werden. So bleibt die Suche nach Belegen bei hochgradig polysemen Stichwörtern eine zeitaufwendige Prozedur, da nur wenige Lesarten eines Stichwortes z.B. einem bestimmten Sachgebiet zuzuordnen sind, sodass Belege zu diesen Lesarten mithilfe entsprechender Korpusauszeichnung auch nicht leicht zu finden wären. Schließlich gibt es auch lexikographische Probleme, die mit den vorhandenen Auszeichnungsmöglichkeiten noch nicht gelöst werden können. Beispielsweise bestünde Bedarf an einer Auszeichnung von festen Verbindungen und idiomatischen Verwendungen, von Datumsangaben oder von wörtlicher Rede.

Die in DeReKo vorhanden Korpusauszeichnungen wurden anhand der sechs Bereiche linguistisches Tagging, bibliographische Metadaten, textlinguistische Klassifikation, textspezifische deskriptiv-statistische Metadaten und Metadaten zu Textduplikaten vorgestellt und auf die unter 2. geschilderten Probleme bezogen. Die Metadaten und Teile des IDS-Textmodells werden automatisch zur Generierung verschiedener Korpus- und Belegansichten in COSMAS II verwendet, und das Wortarten-Tagging ist auch aktiv durch den lexikographischen Nutzer abfragbar. Als Beispiel für die Gewinnung und die automatische Annotation textbezogener textlinguistischer Auszeichnungen wurde die thematische Erschließung der DeReKo-Texte durch einen Naive-Bayes-Klassifikator beschrieben, für

den eine sehr große Menge an Trainingsdaten durch ein Clusteringverfahren gewonnen wurde. Die zur Annotation verwendete Thementaxonomie hat zwar den Anspruch einer gewissen Allgemeingültigkeit, jedoch handelt es sich wie bei den meisten anderen genannten Auszeichnungen um Interpretationen, deren Adäquatheit entscheidend vom jeweiligen Verwendungszweck abhängt. An diesem Beispiel wurde daher auch gezeigt, wie zukünftig benutzerdefinierte Klassifikatoren auf DeReKo anwendbar sein sollen.

Perspektivisch sollen die in COSMAS II z. T. noch bestehenden Diskrepanzen zwischen vorhandenen Auszeichnungen und solchen, die abgefragt werden können, in seinem Nachfolgesystem von vornherein vermieden werden, indem grundsätzlich alle Auszeichnungen für die Abfrage zugänglich gemacht werden. Für eine spätere Ausbaustufe der COSMAS -II-Nachfolgeplattform ist zudem die Möglichkeit vorgesehen, beliebige Annotationen auf der Textebene (wie für Register, Genre oder Thema) und unterhalb der Textebene (wie für POS-Tagging oder syntaktisches Parsing) nutzerseitig hinzuzufügen und sie direkt im Anschluss in der Recherche zu nutzen. Durch diese beiden Strategien soll ermöglicht werden, dass sich mit der Zeit die unter 2. genannten sowie auch weitere auszeichnungsbezogene lexikographische Desiderata auf der Basis von DeReKo umsetzen lassen.

5. Bibliographie

5.1 Wörterbücher

elexiko (2003ff.) = *elexiko*, in: OWID – Online Wortschatz-Informationssystem Deutsch, hg. v. Institut für Deutsche Sprache, Mannheim. Internet: www.owid.de/elexiko_/index.html (10.05.2012).

5.2 Literatur

Abney 1991 = Abney, Steven: Parsing by Chunks. In: Berwick, Robert/Abney, Steven/Tenny, Carol (edd.): Principle-Based Parsing. Dordrecht: Kluwer, 1991, 257–278.

Bański et al. 2012 = Bański, Piotr et al.: The New IDS Corpus Analysis Platform: Challenges and Prospects. In: Calzolari, Nicoletta et al. (edd.): Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), 2012, 2905–2911.

Belica et al. 2011 = Belica, Cyril et al.: The Morphosyntactic Annotation of DeReKo: Interpretation, Opportunities, and Pitfalls. In: Konopka, Marek et al. (edd.): Grammatik und Korpora 2009. Dritte Internationale Konferenz. Mannheim, 22.4.–24.9.2009. Tübingen: Narr, 2011, 451–469.

Dipper 2011 = Dipper, Stephanie: Digitale Korpora in der Lehre – Anwendungsbeispiele aus der Theoretischen Linguistik und der Computerlinguistik. In: Bärenfänger, Maja et al. (edd.): Journal for Language Technology and Computational Linguistics 1. 2011, Themenheft „Sprachressourcen und -technologien in Lehre und Lernen", 81–95.

Erb 2005 = Erb, Sabine: Stichwortansetzung von Partizipien. In: Haß, Ulrike (ed.): Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. Berlin/New York: de Gruyter, 2005, 91–95.

Ide/Bonhomme/Romary 2000 = Ide, Nancy/Bonhomme, Patrice/Romary, Laurent: XCES: An XML-based Standard for Linguistic Corpora. In: Proceedings of the Second Language Resources and Evaluation Conference (LREC), Athen, 2000, 825–830.

Karypis/Zhao 2002 = Karypis, George/Zhao, Ying: Comparison of Agglomerative and Partitional Document Clustering Algorithms. Manual. Minneapolis, MN: Dept. Of Computer Science, University of Minnesota, 2002. Internet: http://users.eecs.northwestern.edu/~yingliu/datamining_papers/paper2.pdf. (10.05.2012).

Keibel/Belica 2007 = Keibel, Holger/Belica, Cyril: CCDB: A Corpus-Linguistic Research and Development Workbench. In: Proceedings of the 4th Corpus Linguistics Conference (CL 2007), Birmingham, 2007. Internet: http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/132Paper.pdf. (10.05.2012).

Klosa 2011 = Klosa, Annette (ed.): *elexiko*. Erfahrungsberichte aus der lexikografischen Praxis eines Internetwörterbuchs. Tübingen: Narr, 2011.

Klosa/Schoolaert 2011 = Klosa, Annette/Schoolaert, Sabine: Die lexikografische Behandlung von Eigennamen in *elexiko*. In: Klosa, Annette (ed.): *elexiko*. Erfahrungsberichte aus der lexikografischen Praxis eines Internetwörterbuchs. Tübingen: Narr, 2011, 193–211.

Klosa/Storjohann 2011 = Klosa, Annette/Storjohann, Petra: Neue Überlegungen und Erfahrungen zu den lexikalischen Mitspielern. In: Klosa, Annette (ed.): *elexiko*. Erfahrungsberichte aus der lexikografischen Praxis eines Internetwörterbuchs. Tübingen: Narr, 2011, 49–80.

Kupietz 2005 = Kupietz, Marc: Near-Duplicate Detection in the IDS Corpora of Written German (Tech. Rep. KT-2006-01). Mannheim: Institut für Deutsche Sprache, 2005.

Kupietz et al. 2010 = Kupietz, Marc et al.: The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In: Calzolari, Nicoletta et al. (edd.): Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), 2010.

Lüngen/Sperberg-McQueen (erscheint) = Lüngen, Harald/Sperberg-McQueen, Michael: A TEI P5 Document Grammar for the IDS Text Model. Erscheint in: Journal of the Text Encoding Initiative.

Möhrs 2011 = Möhrs, Christine: Die typischen Verwendungen in *elexiko*. In: Klosa, Annette (ed.): *elexiko*. Erfahrungsberichte aus der lexikografischen Praxis eines Internetwörterbuchs. Tübingen: Narr, 2011, 81–98.

Perkuhn et al. 2005 = Perkuhn, Rainer et al.: Korpustechnologie am Institut für Deutsche Sprache. In: Schwitalla, Johannes/Wegstein, Werner (edd.): Korpuslinguistik deutsch: synchron – diachron – kontrastiv. Tübingen, 2005, 57–70.

Perkuhn/Keibel/Kupietz 2012 = Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc: Korpuslinguistik. Paderborn: Fink, 2012.

Schiller et al. 1999 = Schiller, Anne et al.: Guidelines für das Tagging deutscher Textcorpora mit STTS. Technischer Bericht. Stuttgart: Institut für maschinelle Sprachverarbeitung, 1999.

Schmid 1994 = Schmid, Helmut: Probabilistic Part-of-Speech Tagging using Decision Trees. In: Proceedings of the International Conference on New Methods in Language Processing. 1994,154–164.

Schnörch 2005 = Schnörch, Ulrich: Die *elexiko*-Stichwortliste. In: Haß, Ulrike (ed.): Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. Berlin/New York: de Gruyter, 2005, 71–90.

Storjohann 2005 = Storjohann, Petra: Das *elexiko*-Korpus. Aufbau und Zusammensetzung. In: Haß, Ulrike (ed.): Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. Berlin/New York: de Gruyter, 2005, 55–70.

Storjohann 2011 = Storjohann, Petra: Paradigmatische Konstruktionen in Theorie, Praxis und im Korpus. In: Klosa, Annette (ed.): *elexiko*. Erfahrungsberichte aus der lexikografischen Praxis eines Internetwörterbuchs. Tübingen: Narr, 2011, 99–129.

Tapanainen/Järvinen 1997 = Tapanainen, Pasi/Järvinen, Timo: A non-projective dependency parser. In: Proceedings of the 5th Conference on Applied Natural Language Processing, Washington DC, 1997, 64–71.

Weiß 2005 = Weiß, Christian: Die thematische Erschließung von Sprachkorpora. Mannheim: Institut für Deutsche Sprache. Online publizierte Arbeiten zur Linguistik (OPAL) 1. 2005.

Witten/Frank/Hall 2011 = Witten, Ian H./Frank, Eibe/Hall, Mark A.: Data Mining. Practical Machine Learning Tools and Techniques. Third Edition. Burlington, MA.: Morgan Kaufmann Publishers, 2011.

5.3 Elektronische Quellen

Belica 1995 = Belica, Cyril: Statistische Kollokationsanalyse und -clustering. Korpuslinguistische Analysemethode. Mannheim, 1995. Internet: http://corpora.ids-mannheim.de/ccdb/. (10.05.2012).

Belica 2001–2007 = Belica, Cyril: Kookurrenzdatenbank CCDB – V 3.2. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. Mannheim, 2002–2007. Internet: http://corpora.ids-mannheim.de/ccdb. (10.05.2012).

Connexor-MPT-Tagset http://www.ids-mannheim.de/cosmas2/projekt/referenz/connexor/ (10.05.2012)

IDS 1991–2012 = Institut für Deutsche Sprache: COSMAS I/II (Corpus Search, Management and Analysis System). Mannheim: Institut für Deutsche Sprache, 1991–2012. Internet: http://www.ids-mannheim.de/cosmas2/. (10.05.2012).

IDS 2012a = Institut für Deutsche Sprache: Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache 2012-I (Release vom 29.02.2012). Mannheim: Institut für Deutsche Sprache, 2012. Internet: http://www.ids-mannheim.de/DeReKo. (10.05.2012).

Netscape 2012 = Netscape. About the Open Directory Project. 2012. Internet: http://www.dmoz.org/about.html. (10.05.2012).

STTS = Stuttgart-Tübingen Tagset. Stuttgart: Institut für maschinelle Sprachverarbeitung. Internet: http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html. (10.05.2012)